HAL
open science

Multi-view Clustering for Hate Speech and Target Community Detection on Social Media

Anaïs Ollagnier, Elena Cabrio, Serena Villata

To cite this version:

Anaïs Ollagnier, Elena Cabrio, Serena Villata. Multi-view Clustering for Hate Speech and Target Community Detection on Social Media. Soph.I.A Summit 2021, Nov 2021, sophia antipolis, France. hal-03466433

HAL Id: hal-03466433

https://hal.science/hal-03466433v1

Submitted on 5 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multi-view Clustering for Hate Speech and Target Community Detection on Social Media

Anaïs Ollagnier, Elena Cabrio, Serena Villata

Université Côte d'Azur, Inria, CNRS, I3S

Introduction

Targeting characteristics of hateful contents through a complete pipeline

Challenges

- Handling with **noisy** and **unstructured** social data (i.e., linguistic errors and idiosyncratic style);
- **Automating** the identification of hate speech relying on the MLMA **multi-aspect hate speech analysis** (i.e., target attributes and target groups).

Vision

- Rethinking the hate speech detection task adopting a clustering approach;
- Extracting hate speech properties reflecting the nature of offensive comments expressed toward target attributes and target groups.

Objectives

Promote → A Multi-view clustering technique to generate fine-grained hate speech target communities.

Explore → The use of multiple data views of a different nature (feature and graph spaces) to improve clustering performance.

Develop → A complete pipeline relying on multilingual pre-trained language models easily adaptable to various social networks.

Multilingual Hate-speech Dataset

The MLMA provides a fine-grained annotation of 5.647 English tweets and 4.014 French tweets. 16 different hate speech target communities are used in this study in French and 26 in English.

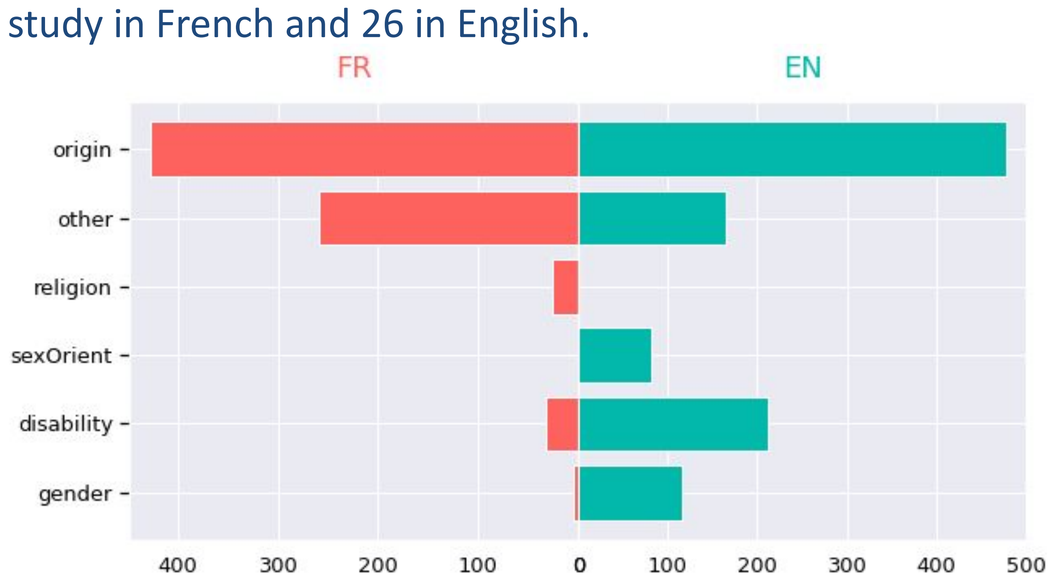

Figure: Distribution of the target attribute in both French and English corpora.

General Architecture

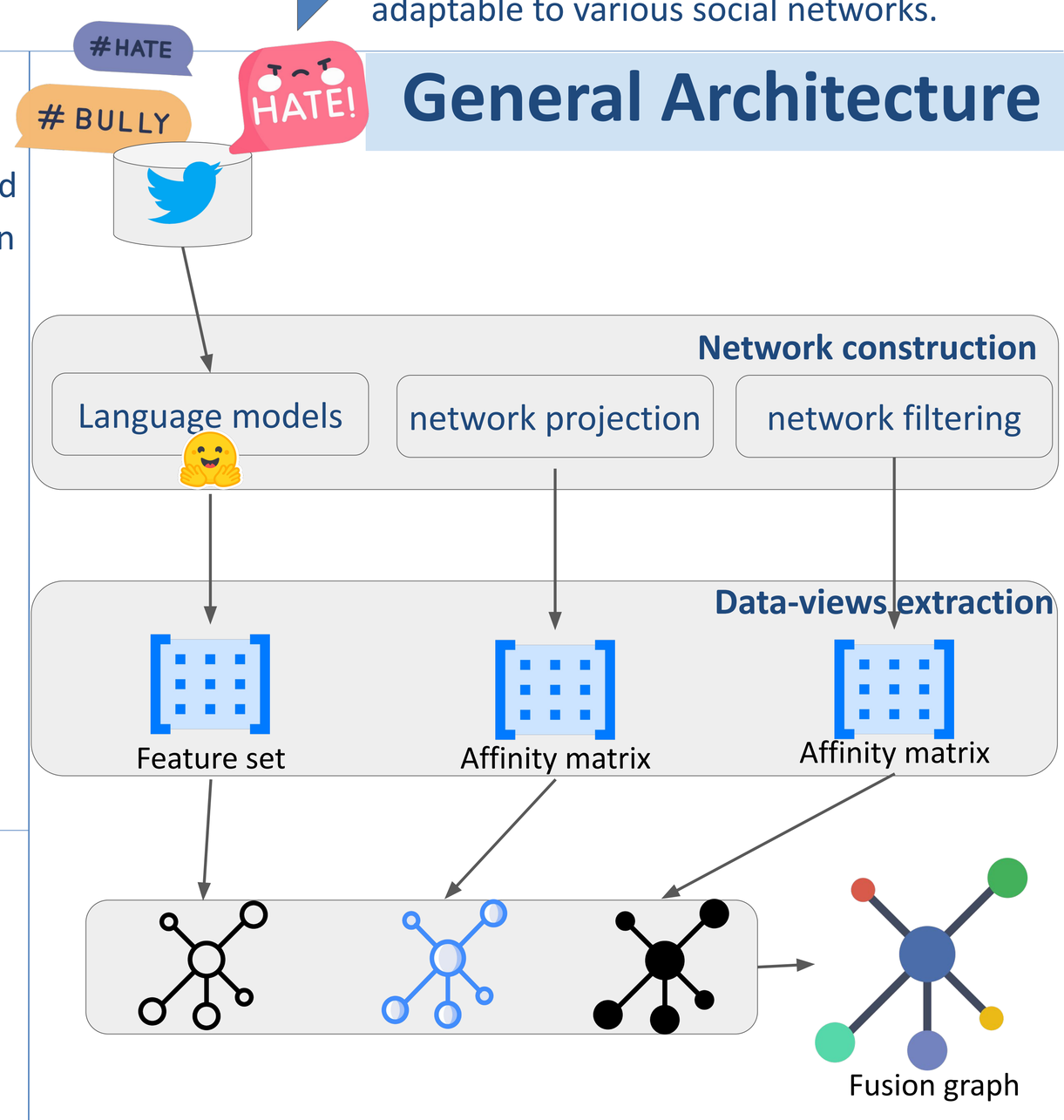

Figure: Multi-view clustering workflow from Twitter data.

Evaluation Results

In total, 69 experiments were conducted:

- **Language models:**
  - mBERT (multilingual BERT)
  - mUSE (multilingual Universal Sentence Encoder)
- **Clustering techniques:** *k*-means, *k*-medoids, spectral clustering and MVSC-CEV
- **Metrics:** Purity, ARI, NMI

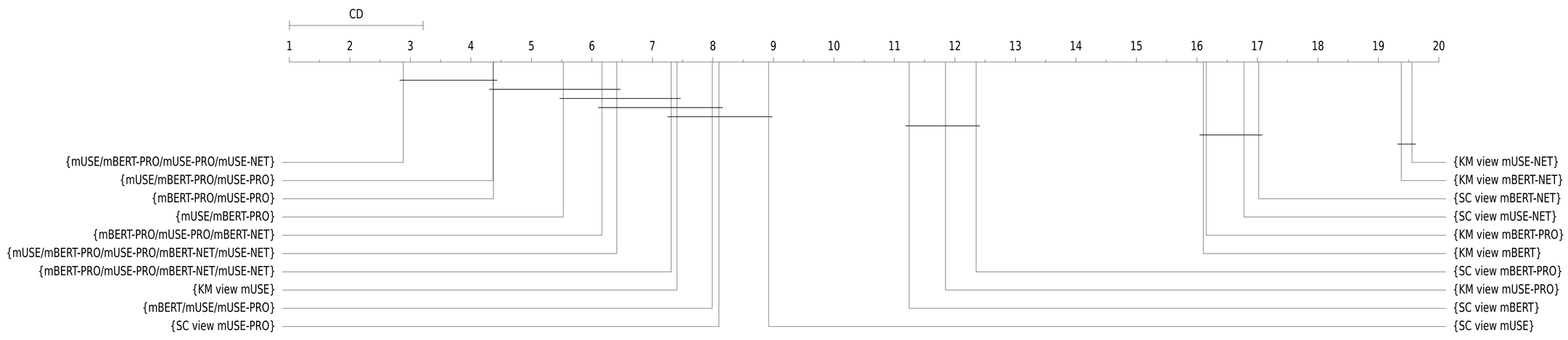

Figure: Models' average ranking resulting from the post hoc Nemenyi test performed on each evaluation metric considering both corpora.

Acknowledgements

This work is funded under the IDEX UCA OTESIA "L'intelligence artificielle au service de la prévention de la cyberviolence, du cyberharcèlement et de la haine en ligne". It has also been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.